International Journal of Innovative Research and Scientific Studies, 8(3) 2025, pages: 5086-5099

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Optimized feature selection based on machine learning models for robust stock market prediction

Mouna Ben Daoud[1], Manel Hamdi[2], Rabeb Younes[3*], Dhouka Oueldoubey[4]

[1]*BESTMOD (Business & Economic STatistics MODeling) Laboratory Tunis Higher Institute of Management, University of Tunis, Tunisia.*
[2]*International Finance Group Tunisia, University of Tunis El Manar, ROMMANA, Tunis Cedex 1068, Tunisia.*
[3]*College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia.*
[4]*National High School of Engineers of Tunis (ENSIT). University of Tunis, Tunisia.*

Corresponding author: Rabeb Younes (*Email: rmyouns@imamu.edu.sa*)

Abstract

This study aims to predict the US financial stock market through machine learning (ML) methods based on optimized feature selection algorithms. Two prediction models were compared: random forest (RF) and support vector regression (SVR). Seventeen variables are used to explain the movement of the S&P 500, NASDAQ, and DJIA indices. These variables are grouped into five categories: basic features, stock market variables, currencies, commodities, and technical indicators. This research work proceeds by applying a variable selection technique to identify the most relevant variables. The optimal set of selected variables was used for forecasting. The results obtained using SVR and RF after variable selection were compared with those obtained before selection. The outcomes of the comparison between these two Artificial Intelligence (AI) methods favor regression after variable selection. Findings show that the feature selection process has a large and significant impact on improving the prediction accuracy of the studied financial markets.

Keywords: Dimensionality reduction, feature selection, random forest, stock market prediction, support vector regression.

DOI: 10.53894/ijirss.v8i3.7708
Funding: This study received no specific financial support.
History: Received: 22 April 2025 / **Revised:** 24 May 2025 / **Accepted:** 28 May 2025 / **Published:** 10 June 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Research on stock market forecasts has gained increasing attention in modern finance. The accurate predictions of stock index movements are crucial for developing effective trading strategies. Therefore, investors can protect themselves against potential market risks while capitalizing on profit opportunities in trading stock indices [1]. Previously, conventional methods included Autoregressive Integrated Moving Average and multivariate regression [2]. Conditional Heteroscedasticity with Generalized Autoregressive [3] has been widely used in forecasting tasks. These traditional methods are unable to provide significant results due to the noisy nature of stock market data. The noisy characteristic refers to the incapacity to fully capture the relationship between future and past prices due to the incomplete information about the historical performance of financial markets. Any information not accounted for in the model is treated as noise. The nonstationary data characteristic indicates that the distribution of financial time series evolves over time. Various factors and unforeseen events, such as economic or political conditions, trader expectations, natural disasters, or wars, lead to changes in financial time series, including stock market indices, exchange rates [4] and oil market indices [5]. Additionally, the abusive use of structured and complex financial instruments can affect financial system stability [6]. Hence, it is crucial to accurately forecast the stock index's directional movements in order to create effective market trading strategies [7].

Researchers are interested in Support Vector Machine (SVM), first suggested by Vapnik and Chapelle [8] to improve forecasting accuracy [9]. Most of the empirical comparison results have shown that SVM outperforms traditional techniques to forecast financial data, especially stock market data [10, 11]. The present study will start by testing the performance of SVR for stock market prediction and then compare it with the Random Forest (RF) model. Many previous works have applied machine learning methods to forecast the stock market [12-15]. Furthermore, the stock market is influenced by a variety of macroeconomic factors, including corporate policies, overall economic conditions, the expectations of investors, and by political events. So, the determination of the most important factors affecting the stock market volatility is needed to enhance the quality of forecasts. This study focuses on the determination of significant explanatory factors for support vector machines. The objective is to choose the most crucial explanatory variables without causing the model's performance to significantly decline. For this reason, an optimized feature selection method was applied.

The current research work aims to provide an appropriate forecasting method for the US financial stock market. Many variables are considered in order to later perform a selection of the most relevant ones, which may have an impact on the evolution of stock market indices. Then, the SVR (Support Vector Regression) and Random Forest methods are applied to forecast the NASDAQ, DJIA, and S&P 500 indices. To identify the suitable method, a predictive performance comparison of these techniques will be conducted.

The remainder of the article is structured as follows: Section 2 presents a literature review related to feature selection. Section 3 defines the supplies and techniques of the study, and Section 4 performs the empirical analysis and discusses the findings. In the end of the manuscript, a general conclusion will be presented.

2. Literature Review Related to Feature Selection

Feature selection methods have been extensively utilized across various disciplines, including medicine, technology, finance, and economics. The main objectives of dimensionality reduction are to mitigate overfitting, improve model performance, and develop faster and more cost-effective models. Dimensionality reduction can be achieved either by selecting relevant features or by extracting them from the dataset. This process changes the original feature representation and generates a new set of features [13]. In addition, feature selection tools preserve the original meaning of the features while selecting the most optimal subset [16].

Actually, feature selection techniques can be divided into three major groups: filter, wrapper, and embedded approaches [17]. These techniques calculate the relevance scores for all features and eliminate those with low scores, keeping the remaining features to be used by a classification algorithm. These filtering techniques rely on the general characteristics of the dataset to select features independently of any classification algorithm. Moreover, these methods are simple and fast in terms of computation, which makes them easily adaptable to large datasets [18]. The wrapper techniques treat the classification algorithm as part of the searching process for the optimal feature subset. The search is conducted by a search algorithm, which acts as a black box, and the optimal subset is integrated with the classification model. However, the embedded techniques incorporate the search for an optimal feature subset directly within the classifier algorithm, where the classifier itself determines the optimal subset of features [16]. In addition, the feature selection involves selecting a subset of the original input variables, typically technical or fundamental indicators. By choosing a relevant subset, the features can more effectively capture the underlying characteristics of the dataset, which can enhance both the accuracy and efficiency of predictions [19]. Many studies have affirmed and validated that feature selection is a key step in modeling stock market prediction [20].

This section examines several feature selection and extraction strategies that have been effectively used in the stock market forecasting scenario. Additionally, the well-combined feature analysis techniques were outlined and assessed.

Żbikowski [21] extended the work of Tay and Cao [22] by modifying the loss function to address SVM regression problems. He proposed that this problem could be reformulated for classification purposes. The study introduced two key innovations. First, it applied a new approach called Volume Weighted (VW)-SVM, which integrates volume information into the loss function to develop a trading strategy. Second, it combined several robust methods, including Fisher's feature selection, VW-SVM, technical indicators, and input vector delays, all incorporated with a walk-forward optimization procedure. The results demonstrated that combining example weighting with feature selection led to significant improvements in trading strategy performance.

Bennasar et al. [23] introduced two innovative nonlinear function selection methods: Joint Mutual Information maximization (JMIM) and Normalized Joint Mutual Information Maximization (NJMIM). These methods leverage mutual information along with the 'maximum of the minimum' criterion to address the issue of overestimating feature significance, as shown through both theoretical and experimental analysis. The proposed methods were evaluated on eleven publicly available datasets, comparing them against five competing methods (CMIM, DISR, mRMR, JMI, and IG). The results revealed that JMIM outperformed the other methods on most of the datasets, reducing the average classification error by nearly 6% compared to the next best method. The statistical significance of these results was confirmed through an ANOVA test. Furthermore, JMIM offered the best balance between accuracy and stability. In the same context, Barak et al. [24] suggested employing Japanese candlesticks for technical analysis and the wrapper Adaptive Neural Fuzzy Inference System-Imperialist Competitive Algorithm ANFIS-ICA in order to forecast stock markets. They developed two approaches for extracting the model's input variables: one based on raw data and another based on signals, with 15 and 24 features respectively. The model outputs buy and sell signals, with prediction accuracy evaluated for periods ranging from 1 to 6 days. In the proposed model, ANFIS predictions are used as the cost function for the wrapper model, while ICA selects the most relevant features. The results showed that the signal-based approach achieved a prediction accuracy of 87%, outperforming the raw-based approach. Additionally, the wrapper feature selection improved predictive performance by 12% compared to the baseline study. Despite being more time-consuming, the wrapper ANFIS-ICA algorithm demonstrated superior time efficiency and higher prediction accuracy when compared to other algorithms like the wrapper genetic algorithm (GA).

In a related study, Su and Cheng [25] introduced a new Adaptive Neuro Fuzzy Inference System (ANFIS) time series model for stock forecasting, which is based on an Integrated Nonlinear Feature Selection (INFS) method. The study integrated a feature selection method to objectively identify key technical indicators. Then, it used ANFIS to create a time series model and test its forecasting performance, strengthening the results with an adaptive expectation model. To assess the model's reliability, the researchers have collected stock market data from the TAIEX and HSI indices between 1998 and 2006 and compared the results with other tools. The proposed model was compared with fuzzy time series models, such as Chen's model [26]. The findings showed that the adopted method outperformed the other models in terms of accuracy, benefit evaluation, and statistical testing.

Pehlivanlı et al. [13] predicted the next day's stock price movement. They used a set feature selection approach to identify the optimal indicator subset. The purpose was to create the best feature subset that accurately predicted future price by eliminating irrelevant and redundant metrics from the data. To achieve this goal, they combined multiple filtering methods such as t-statistics, Fisher score, Relief-F algorithm, and Effective Range-based Gene Selection (ERGS), and then applied SVM for prediction. Finally, they used a voting scheme to integrate the results. Real data records from the Istanbul Stock Exchange (ISE) are considered in their study, including technical and macroeconomic variables. The results showed that applying feature selection improved the forecasting accuracy of stock price direction. Zhong and Enke [14] employed sixty financial and economic characteristics to present a new data mining method to forecast the volatility of the S&P 500 Index ETF (SPY) return. To restructure the data composition, they used three well-known dimensionality reduction approaches on the entire dataset: three types of principal component analysis: Kernel-based Principal Component Analysis (KPCA), Fuzzy Robust Principal Component Analysis (FRPCA), and Principal Component Analysis (PCA). As a result, twelve new datasets were derived from the preprocessed data, each representing different levels of dimensionality reduction. Artificial Neural Networks (ANNs) were then adopted to classify these thirty-six restructured data samples and forecast the daily market return direction. The study also compared the three dimensionality reduction techniques with the original dataset. To validate the results, a series of hypothesis tests were conducted, revealing that combining ANNs with PCA slightly outperformed the two other models in terms of classification accuracy. Additionally, the trading strategies that used projections based on the FRPCA and KPCA models were slightly less profitable than those using PCA and ANNs, and they produced risk-adjusted profits that were much greater than the benchmark methods.

From the previous development, it discloses that a good selection of variables and a suitable choice of optimal variables improve the quality of the forecast in stock market area. Ben Ishak [27] conducted a study comparing SVR and RF to evaluate variable relevance and feature selection. The research provided two main contributions. It performed experimental insights into the effectiveness of variable ranking and selection using both SVR and RF and it established a benchmark to guide researchers in selecting the most suitable method for their data. Experiments on both simulated and real-world datasets showed that the SVR score Gα was more effective for variable ranking in linear cases, whereas the RF score performed better in nonlinear scenarios. Ben Ishak [27] has conducted a comparison between two widely used statistical learning models: SVR and RF. They analyzed data from three monitoring stations in Tunisia to predict the daily maximum ozone concentration (maxO3). These stations covered diverse urban settings, including background, traffic, and industrial areas. The study thoroughly examined the issue of variable selection for regression. The results revealed that RF outperformed SVR in both variable relevance evaluation and variable selection.

In these two previous studies, Ben Ishak [27] and Ishak et al. [28] researchers presented novel approaches to variable selection and applied numerical methods to other sectors; nonetheless, they did not conduct research on the financial market. Therefore, in order to extend their contribution, we attempt to apply the stepwise SVR and RF in the field of finance, more especially in the stock market area.

Rana et al. [29] used a Decision Tree classifier and, Extra Tree classifier to select feature related to the Spanish stock market. They also used Linear Regression (LR), SVR, and Long Short-Term Memory (LSTM) to predict stock market trends. Within the various features, the closing price was selected as the most important using a feature selection algorithm. Additionally, they experimented with varied activation functions and optimizers how they impacted stock price prediction using LSTM. Yuan et al. [30] recently employed a number of feature selection techniques, including the RF model and

Recursive Feature Elimination (RFE), to choose pertinent features. Using eight years data from the Chinese A-share market, they used time-sliding window cross-validation to set the model's parameters. According to their analysis of several integrated models, the best model was achieved when the RF method was applied to both feature selection and stock price movement prediction. Haq et al. [31] created an optimal feature subset by combining characteristics chosen via different feature selection techniques. This optimal feature sample was then utilized in a deep generative method to forecast future price volatility. They calculated an expanded set of 44 technical indicators from the daily stock data of 88 stocks and assessed their relevance by training models separately using logistic regression, SVMs, and RF. The findings showed that integrating features selected by multi-feature selection techniques and feeding them into a deep generative model produces more promising outcomes. Xie and Yu [32] developed an unsupervised feature extraction approach using a Convolutional Autoencoder (CAE) for daily stock market forecasting, which outperformed traditional models. The CAE network integrates convolution and autoencoding techniques for unsupervised feature learning. Their investigations on different stock indices demonstrated a significant improvement in prediction accuracy compared to standard methods.

Recently, a time-efficient Hybrid Stock Trends Prediction Framework (HSTPF) was presented by Bhanja and Das [33] with the aim of accurately predicting stock market trends, especially during Black Swan events. They evaluated the effectiveness of many machine learning classifiers and used Black Swan event analysis and feature selection to improve the prediction accuracy of HSTPF. They revealed that the framework is effective in terms of computational time and beat comparable methods in terms of prediction accuracy, especially during Black Swan events.

To the best of our knowledge, the current research study presents an early attempt to incorporate features selection algorithms with machine learning models into stock market fluctuation prediction. Moreover, this study combines SVR criteria with stepwise search space algorithms for finance feature selection area.

3. Supplies and Techniques

This section presents techniques used for modeling. Financial variables were processed and fed into two machine learning algorithms: Support Vector Regression (SVR) and Random Forest (RF). Feature selection was applied using stepwise and backward search space algorithms to optimize predictive performance. Both techniques are based on artificial intelligence, and they have also been widely used in various fields, including finance. These methods are grounded in statistical learning theory; thus, they have a strong theoretical foundation. Moreover, in practice, they consistently show better predictive performance compared to traditional techniques.

3.1. Support Vector Regression (SVR)

Proposed by Vapnik [34] SVM has become one of the most robust machine learning models for forecasting [35]. SVMs are prominent for their foundation in the conventional approach of empirical risk minimization (ERM), which has been demonstrated to be less successful than structural risk minimization (SRM) [36]. While ERM focuses on reducing errors based on the training data, SRM attends to minimize the overall expected risk. SVMs not only focus on fitting the model to the data but also on how well it generalizes to new and unseen data. Over the time, researchers have shown that SVMs can significantly improve predictions when new data is involved. Originally designed for classification tasks, SVMs have now been adapted to handle regression problems as well.

3.1.1. Model presentation

This section presents an overview of the fundamental concept and formulation of SVR. There are two types of models for SVM: SVM for classification and SVM for regression called SVR. The analysis focus on the regression context using the classical SVR model introduced by Vapnik [37]. The goal of the SVR is to estimate the best function:

$$f(x) = \left(w, \Phi(x)\right)_H + b \qquad (1)$$

Where, w is weight vector, b is the bias parameter and $\Phi(x)$ is the nonlinear mapping from input space to high dimensional feature space. The function f(x) approximates best the relationship between the input vector x_i and the output vector y_i.

Therefore, the regularized regression risk is minimized, as described in Vapnik [37] and Smola et al. [35]. The regularization function R is presented as follows:

$$R[f] = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} L(x_i, y_i, f) \qquad (2)$$

The approximation quality is measured by the loss function L:

$$L(x_i, y_i, f) = | y - f(x) |_\varepsilon^2 \qquad (3)$$

the following optimization issue must be resolved using the SVR model:

$$\text{minimize } \| w \|^2 + C \sum_{i=1}^{l} \left(\xi_i^2 + \hat{\xi}_i^2\right) \qquad (4)$$

$$\text{subject to} \begin{cases} ((w.x_i) + b) - y_i \leq \varepsilon + \xi_i, \\ y_i - ((w.x_i) + b) \leq \varepsilon + \hat{\xi}_i, \\ \xi_i, \hat{\xi}_i \geq 0, i = 1,2 \ldots l \end{cases}$$

For further details, refer to Vapnik [37].

The capacity of the SVM model to use the kernel function to solve the data linear indivisibility problem is one of the most significant advantages. Data in a low-dimensional space are known to be nonlinear, but they can become linearly separable when mapped to a high-dimensional space. The kernel function reduces complexity and facilitates the transition

from low to high dimensions by calculating the inner product of two vectors in the low-dimensional space and mapping it to the high-dimensional space. Although there are different kinds of kernel functions, the most widely used are listed below:
Linear: $K(x_i, x_j) = x_i \,.\, x_j$
Polynomial: $K(x_i, x_j) = (1 + x_i \,.\, x_j)^{\rho}, \rho > 0$
Gaussian:$K(x_i, x_j) = \exp(-\| x_i - x_j \|/\sigma^2)$.

3.1.2. Variable Selection Criteria Based on SVR

This study uses SVR bounds and some components of bounds like ranking criteria, because the variable selection process requires a ranking criterion to rank variables. These criteria are introduced and explained through 4 functions:

$$G_R(\alpha, \hat{\alpha}) = \widetilde{R}^2 \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i) \quad (5)$$
$$G_S(\alpha, \hat{\alpha}) = \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i)\, \tilde{S}_i^2 \quad (6)$$
$$G_\alpha(\alpha, \hat{\alpha}) = \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i) \quad (7)$$
$$G_W(\alpha, \hat{\alpha}) = \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i)\, (\hat{\alpha}_j + \alpha_j) K(x_i, x_j) \quad (8)$$

$G_R G_S$are criteria are based on SVR bounds. These criteria are used in the selection of relevant variables. The most relevant is the one that tends to decrease the bound when used and is the top-ranked variable. It is important to note that α is not a bound itself, but rather a term that appears in both radius-margin and span-estimate bounds. Therefore, minimizing this term will also minimize those bounds. These criteria are directly linked to the performance of the SVR. More details are presented by Rakotomamonjy [38].

3.2. Random Forests (RF)

Random Forests (RF) are widely used and highly efficient algorithms that rely on model aggregation techniques, applicable for both classification and regression tasks. Breiman [39] they have attracted the attention of many researchers due to their performance and robustness in forecasting and variable selection. The current study focuses on their use for solving regression problems.

3.2.1. Model presentation

Random forests are a modeling approach that generates estimators for either the Bayes classifier, which aims to minimize the classification error, or for the regression function.

Assuming that B is independent, identically distributed (i.i.d.) random variables, each with variance σ^2, are averaged. Their average variance is $(1/B)\sigma^2$. However, if the variables have a positive pairwise correlation ρ and are identically distributed but not necessarily independent, the average's variance is as follows:

$$\rho\sigma^2 + ((1-\rho)/B)\sigma^2. \quad (9)$$

The second term tends to be negligible as B increases. Consequently, the advantages of averaging are constrained by the correlation between tree pairs throughout the bagging process. In order to prevent the excessive variance rising, random forests aim to improve variance reduction by bagging while reducing the connection between trees. During the tree-growing process, input variables are chosen randomly. The steps involved in creating a tree with a bootstrapped dataset are as follows: Choose m≤p of the input variables randomly to be split candidates prior to each split, where m values are usually as low as 1 or equal to √p. Following the growth of B such trees $\{T(x, \theta_b\}_1^B$, the formula below displays the RF predictor for regression:

$$\hat{f}_{rf}^{B} = \frac{1}{B} \sum_{b=1}^{b} T(x;\, \theta_b) \quad (10)$$

3.2.2. Feature Selection Based on RF

The objective of the feature selection method based on RF is to order variables based on predetermined standards, such as significance metrics. While the second measure takes into account the average decrease in node impurity, the first measure assesses importance by calculating the average increase in prediction error. Breiman [40] presented a commonly used relevance score in the context of random forests for regression, which is the increase in Mean Squared Error (MSE) that occurs when the values of a particular variable are randomly permuted inside the Out-Of-Bag (OOB) samples. The RF predictive accuracy can be measured using the Out-Of-Bag (OOB) data as follows:

$$OOB_{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\hat{y}}_{OOBi})^2 \quad (11)$$

$\bar{\hat{y}}_{OOBi}$ represents the average prediction for the ith observation across all trees where this observation was part of the Out-Of-Bag (OOB) sample.

The importance score of variable j in a random forest is calculated as the average across all ntree trees:

$$G_{RFj} = \frac{1}{ntree} \sum_{t=1}^{ntree} (\widetilde{OOB}^{j}_{MSE} - OOB^{t}_{MSE}) \quad (12)$$

3.3. Search-Space Algorithms

Search-space algorithms are optimization techniques used to efficiently explore possible combinations of features in a dataset [41]. The variable selection process is carried out in two stages: First, all variables are ranked in descending order of relevance based on the SVR criteria and RF. Then, stepwise and backward forward and backward algorithms are applied to identify the subset of variables that best explain the data.

3.3.1. Backward Algorithm

The greedy iterative algorithm based on backward selection or elimination has been used in many applications [38]. The backward feature elimination method introduced by Guyon et al. [42] is basically a recursive process that ranks features according to some measure of their importance in minimizing the criterion. During each iteration, feature importance is assessed, and the less relevant ones are eliminated. An alternative approach, not used here, serves to remove a group of features each time in order to speed up the process [43]. This backward selection method consists of considering initially all variables and then eliminating one of them at a time. In this stage, the variable that increases the leave-one-out bound is removed, and this variable is last ranked. However, the variable that decreases the leave-one-out bound is the last removed and is the top-ranked.

3.3.2. Stepwise Algorithm

This study adopts a stepwise forward approach, inspired by the sequential variable introduction method used by Ghattas and Ben Ishak [44]. Initially, a series of progressively more complex models are constructed by incorporating the (k) most important variables, with one variable added at each step. When the number of variables (p) becomes large, the process is adjusted, and additional variables are introduced in blocks to manage the complexity. Then, the error rate of each model is estimated using stratified random splitting for the SVR model and evaluated on OOB samples for the RF. The variables obtained in the model with the lowest error rate are chosen. Unlike the search-space approaches based on the backward algorithm, this algorithm allows for the automatic identification of the chosen subset size of relevant predictors.

4. Empirical Analysis

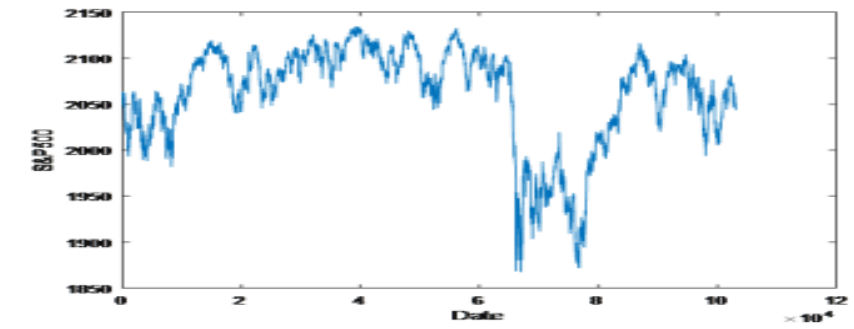
The effectiveness of the RF and SVR methods are compared in forecasting and modeling stock markets prices. The Mean Square Error (MSE) and Mean Absolute Error (MAE) are the performance indicators used to assess the forecasting capabilities of the suggested methods.

4.1. Data Collection and Input Selection

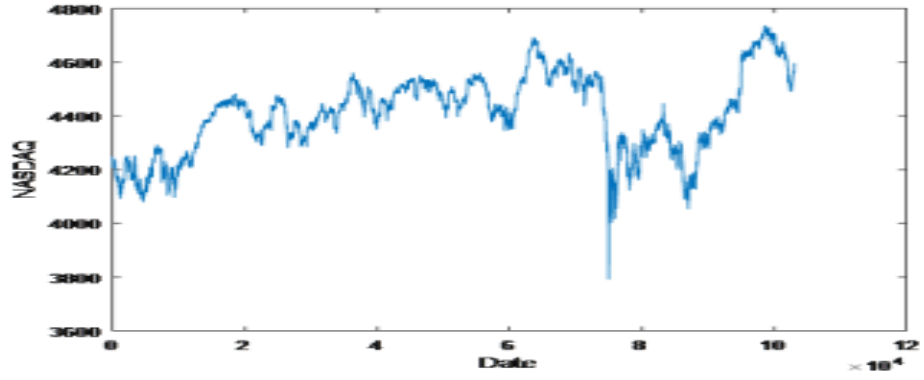
In the present study, the American stock market was investigated. Different numerical applications of methods presented in this work concern the indices DJIA, S&P 500, and NASDAQ at a daily frequency. The data were collected from Thomson Reuters DataStream. Each of the three databases contains seventeen explanatory variables. The observation period is from 05 January 2011 to 04 October 2021, accumulating 2705 observations.

Before starting the empirical analysis, the basis statistics of the data used were studied.

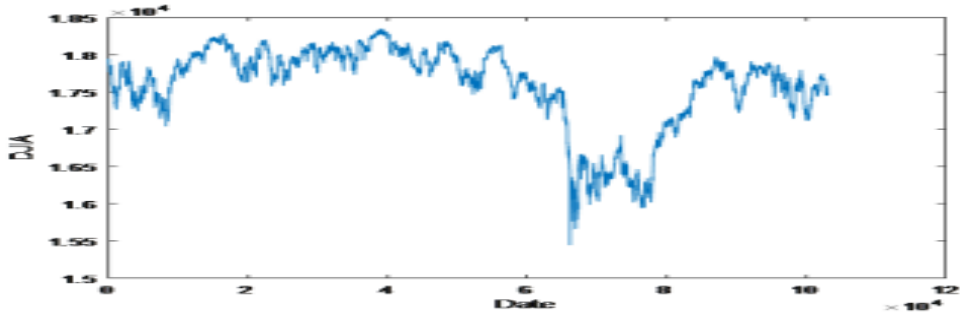
Figure 1 shows the DJIA, S&P500, NASDAQ indexes daily averages during observation period.



(a)



(b)



(c)

Figure 1.
(a) S&P 500 index daily average, (b) NASDAQ index daily average, (c) DJIA index daily average.

It is important to consider the level of volatility. This concept informs about the situation of the financial markets: high volatility reflects a context of investor uncertainty regarding the financial markets. The investment in a stock market needs to be vigilant about the variations of its index.

Figure 1 indicates that the concerned indices (S&P500, NASDAQ, and DJIA) are volatile and variable over time. Some investors are interested in volatile assets; others prefer less risky assets with low volatility. Hence, the study aims to provide effective forecasting to guide traders to efficient trading strategies and appropriate decisions.

Table 1.
Descriptive analysis of daily databases.

	Min	Median	Mean	Max	SD
S&P 500	735.09	1377.94	1432.52	2107.39	340.920955
NASDAQ	1377.84	2620.34	2951.999835	5128.28	975.2838674
DJIA	7062.93	12569.79	12917.38	18132.70	2701.13628

Table 1 presents a summary of basic statistics. In this work, seventeen variables are used to explain daily DJIA, S&P500 and NASDAQ stock indexes. These variables are grouped into five categories: basic features, stock market variables, currencies, commodities, and technical indicators. A crucial step in the development of a forecast model is the choice of input variables. The choice is based on previous studies [45-48]. The explanatory variables are presented in Table 2. The most relevant variables are selected hereafter.

Table 2.
Explanatory variables.

Categories	Variable	Definition
Commodities	WHE	wheat price
	COFA	coffee price
	GAZ	natural Gaz price
	OIL	oil price
	COP	copper price
	GOLD	gold price
Currencies	EUR/USD	Euro/Dollar exchange rate
	USD/JPY	Dollar/Japanese Yen exchange rate
	GBP/USD	Great Britain pounds/Dollar exchange rate
Stock market Variables	SP	S&P 500 daily price
	NQ	NASDAQ daily price
	DJ	DJIA daily price
Basic features	O	Opening index values
	H	High indexe values
	L	Low index values
Technical indicators	HC	H/C
	OC	O/C
	WHE	L/C

4.2. Findings and Interpretation

In the current study, all data are normalized before being used in training models. They are normalized to the zero-mean and unit variance in order to be optimally processed by the model and enhance forecasting accuracy [49].

Employing SVR, the kernel function is first determined. As a kernel function, either a Gaussian or a polynomial kernel is used. However, the Gaussian kernel resulted in excessively long computation times, making it impractical for our study. Therefore, in this research, a polynomial kernel with different degrees is opted.

Before performing the forecasting task with the SVR model, key hyperparameters must firstly be determined: the error cost C, the width of the tube ε-tube, and the degree of the polynomial kernel d. The Grid Search method is used on the training dataset in order to identify these parameters. Various combinations of (C, ε, d) are tested, and the one that minimizes the Mean Squared Error (MSE) is chosen. Empirical results indicate that the optimal parameters (C, ε; d) are respectively (1,0.01,1) for S&P500, (100,0.01,1) for NASDAQ and (100,0.001,1) for DJIA.

The primary parameters of a random forest model are mtry, which controls the number of input variables randomly selected at each split, and ntree, which determines the total number of trees in the forest. Additionally, a third parameter called node size specifies the minimum size for the leaves of the trees.

- The prediction of particular target point x is determined by averaging the output from each individual tree, when using random forests for regression tasks. In addition, the inventors make the following suggestions:
- The minimum node size for classification is 1, and the default value for mtry is [√p].
- The minimum node size for regression is 5, and the default value for mtry is [p/3].

Generally, the default value 5 is retained for all experiments, since it is close to the maximal tree choice (ntree).

The SVR and RF models are evaluated using various training and testing data size ratios: 90% to 10%, 80% to 20%, and 75% to 25%, to identify the optimal data split. Results are presented in Table 3.

Table 3.
Results of train/test selection.

		mean MSE		
		75% vs. 25%	**80% vs. 20%**	**90% vs. 10%**
SVR	S&P500	65.12	67.86	69.33
	NASDAQ	43.30	45.87	49.85
	DJIA	82.32	85.63	89.65
RF	S&P500	25.66	26.51	28.33
	NASDAQ	31.20	32.87	35.47
	DJIA	37.24	38.96	42.63

The experimental results reveal that the 75% to 25% data split was consistently the most effective, yielding to the lowest mean MSE for all three datasets.

Different Wrapper algorithms are employed in the experiment. These algorithms have used nine variables ranker algorithms. Eight of them are based on SVR criteria (G_R, G_S, G_W, G_α) introduced by Rakotomamonjy [38]. In the actual study they will be combined with both backward and stepwise algorithms. The Combination of stepwise with these criteria was introduced by Ben Ishak [27]. Considering the variable selection methods related to backward algorithms, the top 5 ranked variables have been utilized as predictive inputs for y. In the following study five databases are obtained; the first uses all variables, the second uses five top-ranked variables according to G_α, the third uses five top ranked variables according to their margin G_W , the fourth uses five top-ranked variables according to their radius margin G_R and the last one uses five top-ranked variables according to span estimate bound G_S, as a criterion. Considering variable selection methods based on stepwise algorithm, the technique provides us with the optimal package of important variables and eventually gets four databases. One uses the optimal packet of variables combining stepwise and G_α, one uses stepwise with G_W, one uses stepwise with $G_R$and the last one uses stepwise algorithm with G_S criterion. In addition, stepwise is combined with Random Forest and also generates a new database with an optimal packet of variables. These results are compared with those of Random Forest using all variables.

Hence, it can be deduced that in the actual study, results of eleven new databases will be compared for each index that gives a total of 33 databases.

Political developments, corporate policies, economic conditions, commodities price, macroeconomic factors, investor expectations, institutional investor decisions, even investor psychology and affect stock market fluctuations. Additionally, a variety of technical criteria are employed to extract statistical data from stock price values.

This study aims also to predict the stock prices direction by using an optimal set of indicators selected through an ensemble of feature selection strategy. The objective is to identify the most important subset of features that enhances prediction accuracy while eliminating unnecessary and redundant variables from the dataset. Consequently, wrapper methods based on SVM and RF have been carried. This study can be part of studies dealing with the problem of variables selection and reduction of explanatory variables dimensionality. In this investigation, four criteria for variable selection resulting from SVR model combined with stepwise and backward feature elimination algorithms are used. Then, stepwise technique is combined with random forest. SVR and RF after selection of variables are compared with SVR and RF before variable selection. So, for each index eleven databases are compared. A variety of variables are utilized: financial, stock market and technical variables. The choice of technical variables is based on the work of Barak et al. [24]. The selection marked the following variables as relevant: Open, High, Low, LC, EUR/USD. Results of selection are summarized in Appendix A.

After applying feature selection, forecasting findings of SVR and RF algorithms combined with feature selection are compared with those without feature selection. For each train, the results were averaged over 20 trials with random sampling of the training data. The mean MSE and MAE values obtained from these 20 trials are presented in the Tables 4, 5, and 6.

Table 4.
S&P500 Forecasting results.

Forecasting model	Selection algorithm	MSE	MAE
SVR	All	28.39299227	4.239713183
	SAlpha	1.33E-06	0.000945849
	SMargin	3.01E-05	0.0039455
	SRadius	1.33E-06	0.000945849
	SSpan	0.0058	0.0558
	BAlpha	0.0173	0.0815
	BMargin	0.0477	0.1688
	BRadius	0.042817394	0.150687279
	BSpan	0.1607	0.3010
RF	All	0.245938202	0.298494224
	SRF	0.244121669	0.29774504

Table 5.
NASDAQ Forecasting results.

Forecasting model	Selection algorithm	MSE	MAE
SVR	All	69.85655475	6.352712264
	SAlpha	2.39E-06	0.001298903
	SMargin	9.91E-05	0.006571624
	SRadius	3.93E-05	0.004986682
	SSpan	0.267330228	0.349791948
	BAlpha	0.626387318	0.616387318
	BMargin	0.854773085	0.69670882
	BRadius	0.8097	0.6184
	BSpan	0.8769	0.7282
RF	All	3.0724	0.8290
	SRF	1.9330	0.6155

Table 6.
DJIA Forecasting results.

Forecasting model	Selection algorithm	MSE	MAE
SVR	All	77.239289	30.04174939
	SAlpha	1.73E-05	0.0033835
	SMargin	0.0058	0.0558
	SRadius	0.001921706	0.032411072
	SSpan	0.0173	0.0815
	BAlpha	0.0477	0.1688
	BMargin	3.253529004	1.340287794
	BRadius	0.267330228	0.349791948
	BSpan	3.509603582	1.389270951
RF	All	5.244308504	2.547064601
	SRF	5.266413304	2.545803192

Predicting stock market movements is a significant challenge in both finance and economy. Historically, one of the main approaches of forecasting stock prices relied solely on past data. However, other factors such as stock exchange fluctuation, commodities prices, economic environment and many other indicators can impact stock markets.

Stock market investors aim to maximize their profits, which requires effective tools to analyze stock prices and market trends. Machine learning algorithms have been established to create advanced prediction models that can accurately forecast stock prices and identify market trends. Several models have been proposed to account for the various factors influencing stock prices. This research work specifically focuses on the adoption of machine learning algorithms, such as SVM and RF, to improve prediction accuracy.

The results of the proposed two models, SVR and RF are compared. For each model, the predictions are completed using seventeen explanatory variables related to three indices SP&500, NASDAQ and DJIA. Moreover, the prediction performance is evaluated using MSE and MAE. Results are summarized in Table 4. SVR and RF are two approaches derived from statistical learning theory, providing them with a solid and robust theoretical basis. Also, on the operational level, they have shown great robustness compared to various existing forecasting models.

The Tables 4, 5, and 6 show that SVR criteria combined with Stepwise Search space algorithm outperform SVR combined with Backward search space for each index. Additionally, it is clear that the criterion based on alpha Gα outperforms other criteria for both stepwise and backward algorithms. Also, GW outperforms GR and GR outperforms GS for each index. Furthermore, results obtained after feature selection based on Random Forest combined with Stepwise

outperform results of forecasting using random forest without feature selection. So, the combination of stepwise with random forest may improve forecasting results. Considering only forecasting without feature selection, it is clearly seen that Random Forest gives better forecasting results than SVR but combined with feature Selection SVR algorithm exceeds Random Forest.

To be consistent with the literature, a comparison between the current work and other recent studies was carried out. To this end, five recent studies based on feature selection methods were analyzed and summarized in Table 7.

Table 7.
Findings comparison with literature related to feature selection incorporating in stock market prediction.

Authors	Features	Techniques for feature selection	Forecasting methods	Datasets and periods
Rana, et al. [29]	Fundamental characteristics.	Decision tree and extra tree classifiers.	SVR, LR, LSTM	Spanish stock market (from 1-1-2008 to 31-12-2018)
Yuan, et al. [30]	Fundamental and technical indicators.	SVM-RFE, RF	SVM, RF, ANN	Chinese A-share stocks (from 1-1-2010 to 1-1-2018)
Haq, et al. [31]	Fundamental characteristics and technical indicators.	LR, SVM, RF	Deep generative model	88 stocks from NASDAQ (from 01-01-2014 to 01-01-2016)
Xie and Yu [32]	Commodities, the U.S. dollar's exchange rate to other currencies, technical indicators, international stock market indices, and data from big businesses.	Autoencoder	SVM	SPY, NASDAQ, HSI, DJIA, and SSEC indices (from 01-01-2010 to 31-12-2019)
Bhanja and Das [33]	Technical indicators	Autoencoder	DLM, MNB, SVM, KNN, AB, GBM	S&P BSE SENSEX (from 01-01-1991 to 31-03-2021) and Nifty 50 (from 01-01-1996 to 31-03-2021) indices
Present work	Basic features and Stock market, variables, commodities, currencies, technical indicators	SVM-RFE, SVM-Stepwise, RF	SVM, RF	NASDAQ, DJIA, S&P500 (from 01-01-2010 to 31-12-2019)

Rana et al. [29] adopted for variable selection two embedded methods: Decision Tree Classifier and Extra Trees. This means that there is no direct relationship between variable selection and the forecasting process. In contrast, here, a wrapper approach is adopted where the variable selection process and the forecasting procedure are interdependent. The application of variable selection in their work highlighted the superiority of the closing value compared to other baseline variables such as Opening price, Low price, High price and, Volume. However, in the current study, these variables (Opening price, Low price, High price) were selected among the most relevant variables. Considering the forecasting effort in Rana et al. [29] it is certain that LSTM can provide better predictions than linear regression (LR) and support vector regression (SVR).

Yuan et al. [30] used both Recursive Feature Elimination (RFE) for SVM and feature selection based on Random Forest (RF) for the variable extraction procedure. However, in the present work, for SVM-based variable selection, two search space algorithms are employed: the Recursive Feature Elimination (RFE) method, already used by Yuan et al. [30], and additionally, the Stepwise selection technique is incorporated. In the actual study's variable selection process, for the Stepwise-RF and Stepwise-SVR combinations, the optimal packet of selected variables is considered. However, when using the SVM-RFE procedure, the top 5 ranked variables were retained. Whereas in Yuan et al. [30], the top 80% of all features are selected, which means that there are 48 selected features. Additionally, in the same work, the best stock price trend forecasting results were obtained when selecting features using Random Forest (RF) and applying RF for stock price trend prediction. Whereas, in the actual analysis, when considering forecasting without feature selection, Random Forest clearly outperforms Support Vector Regression (SVR), and, when combined with feature selection, the SVR algorithm surpasses Random Forest in forecasting performance.

Haq et al. [31] utilized feature-based and technical indicators for stock market forecasting. However, five categories of variables were employed in our study. They adopted an embedded approach for feature selection, ranking variables using L1–LR, SVM, and RF, and applied the MFFS technique, which combines these three feature selection methods. As a result, they obtained four different subsets during the feature selection stage. However, our work employs wrapper feature selection methods based on SVR and RF. Haq et al. [31] found that variable selection using RF outperformed SVM. Indifference, our study revealed that SVM surpassed RF in variable selection. Additionally, the authors proposed a new selection technique combining SVM, RF, and L1–LR, which achieved the best outcomes when compared to the individual performance of the three selection techniques.

Xie and Yu [32] chose 95 financial and economic indicators in order to anticipate the stock market. These factors are divided into five categories: technical indicators, global stock market indexes and U.S. exchange rates, the dollar to other currencies, commodities, and data from major companies. Their choice of variables is closely similar to our factor choice, which classifies the variables into five categories: commodities, currencies, stock market variables, basic features, and technical indicators. They used a feature extraction method based on a CAE network designed by combining convolutional and autoencoders. Although in our study, a wrapper feature selection method based on RF and SVR is applied and combined with stepwise and RFE search space algorithms. They assumed that the average accuracy of the CAE technique was approximately 3% higher than other models (i.e., LSTM, DNN, SVM, and PCA) for the five studied stock indices.

Bhanja and Das [33] adopted only technical indicators in their study and used an autoencoder for feature extraction. They compared the forecasting performance of six methods, a DLM and five machine learning methods like MNB, SVM, KNN, AB, and GBM. Results revealed that the proposed hybrid stock trends forecasting framework (HSTPF) outperforms other existing studies. Whereas, in the actual study, wrapper methods based on SVR and RF are used, and the forecasting performance of those two methods is compared.

5. Conclusions

Predicting the stock market is a challenging task for investors and researchers, primarily due to its inherent complexity and the multitude of unpredictable factors involved. The task requires sophisticated analysis and an understanding of various market dynamics. As a result, highly accurate forecasting models have become essential tools, often deployed in automated trading systems. This research focuses on the study of the American stock market, specifically the analysis of the S&P 500, NASDAQ, and DJIA indices, which represent the primary indices of the American stock market. Several technical indicators are selected as model inputs, such as currencies, commodities, and technical indicators.

The experiments were crucial, as they not only allow to predict the stock market volatility but also provide valuable insights into the nature of the data. These insights are helpful in improving the training of the classifiers in the future.

The actual study is based on the comparison between two algorithms: SVR and RF, and the optimization of feature selection. By considering the forecasting without feature selection, it is clearly seen that Random Forest gives better forecasting results than SVR, but if combined with feature selection, the SVR algorithm exceeds Random Forest.

Therefore, the findings presented a complement to previous works and confirm the existence of interaction between the different stock markets and a powerful relationship between currency market and stock market. The introduction of technical indicators appears to be important in daily frequency. Also, introducing new variables to the historical values of the stock market provides an improvement in forecasting results.

The use of the proposed prediction models could be extended to various other fields, such as forecasting GDP, predicting energy consumption trends, estimating commodity prices, or even weather forecasting. The potential of these models, when adapted to different domains, can be explore to provide accurate, data-driven insights across a wide range of industries, helping to improve decision-making. Several other factors can significantly impact the stock market, including people sentiment, news events, and developments both within the country and globally.

Since both SVR and RF techniques are robust and can handle high-dimensional data, the framework of the present investigation can be easily extended to incorporate additional relevant input variables. It would be interesting to compare these algorithms with others like Naïve Bayes and some deep learning algorithms [50, 51]. For further research development, other methods of variable selection can be introduced and using high-frequency data to improve stock market modeling.

References

[1] K. Manish and M. Thenmozhi, "Forecasting stock index movement: A comparison of support vector machines and random forest," in *Proceedings of Ninth Indian Institute of Capital Markets Conference, Mumbai, India. http://ssrn.com/abstract=876544*, 2005.

[2] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307-327, 1986. https://doi.org/10.1016/0304-4076(86)90063-1

[3] G. E. Box and G. M. Jenkins, *Time series analysis, forecasting, and control*. San Francisco: Holden-Day, 1970.

[4] M. Qureshi, N. Ahmad, S. Ullah, and A. R. ul Mustafa, "Forecasting real exchange rate (REER) using artificial intelligence and time series models," *Heliyon*, vol. 9, no. 5, 2023. https://doi.org/10.1016/j.heliyon.2023.e16335

[5] M. Hamdi and C. Aloui, "Forecasting crude oil price using artificial neural networks: a literature survey," *Econ. Bull*, vol. 35, no. 2, pp. 1339-1359, 2015. https://doi.org/10.1016/j.procs.2018.05.050

[6] R. Younes, "Investigation on the credit risk transfer effects on the banking stability and performance," *Cogent Economics & Finance*, vol. 10, no. 1, p. 2085264, 2022. https://doi.org/10.1080/23322039.2022.2085264

[7] M. T. Leung, H. Daouk, and A.-S. Chen, "Forecasting stock indices: A comparison of classification and level estimation models," *International Journal of Forecasting*, vol. 16, no. 2, pp. 173-190, 2000. https://doi.org/10.1016/s0169-2070(99)00048-5

[8] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Computation*, vol. 12, no. 9, pp. 2013-2036, 2000. https://doi.org/10.1162/089976600300015042

[9] C.-Y. Yeh, C.-W. Huang, and S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2177-2186, 2011. https://doi.org/10.1016/j.eswa.2010.08.004

[10] L.-J. Kao, C.-C. Chiu, C.-J. Lu, and C.-H. Chang, "A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting," *Decision Support Systems*, vol. 54, no. 3, pp. 1228-1244, 2013. https://doi.org/10.1016/j.dss.2012.11.012

[11] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, "Stock index forecasting based on a hybrid model," *Omega*, vol. 40, no. 6, pp. 758-766, 2012. https://doi.org/10.1016/j.omega.2011.07.008

[12] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," *arXiv preprint arXiv:1605.00003*, 2016.

[13] A. Ç. Pehlivanlı, B. Aşıkgil, and G. Gülay, "Indicator selection with committee decision of filter methods for stock market price trend in ISE," *Applied Soft Computing*, vol. 49, pp. 792-800, 2016. https://doi.org/10.1016/j.asoc.2016.09.004

[14] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," *Expert Systems with Applications*, vol. 67, pp. 126-139, 2017. https://doi.org/10.1016/j.eswa.2005.06.024

[15] K. Iqbal *et al.*, "Forecasting stock market using machine learning approach encoder-decoder convlstm," presented at the International Conference on Frontiers of Information Technology (FIT), 2021.

[16] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007. https://doi.org/10.1093/bioinformatics/btm344

[17] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & electrical engineering*, vol. 40, no. 1, pp. 16-28, 2014. https://doi.org/10.1016/j.compeleceng.2013.11.024

[18] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, *Feature extraction: Foundations and applications*. Germany: Springer, 2006.

[19] L.-P. Ni, Z.-W. Ni, and Y.-Z. Gao, "Stock trend prediction based on fractal feature selection and support vector machine," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5569-5576, 2011. https://doi.org/10.1016/j.eswa.2010.10.079

[20] C.-F. Tsai and Y.-C. Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches," *Decision support systems*, vol. 50, no. 1, pp. 258-269, 2010. https://doi.org/10.1016/j.dss.2010.08.028

[21] K. Żbikowski, "Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1797-1805, 2015. https://doi.org/10.1016/j.eswa.2014.10.001

[22] F. E. Tay and L. Cao, "Modified support vector machines in financial time series forecasting," *Neurocomputing*, vol. 48, no. 1-4, pp. 847-861, 2002. https://doi.org/10.1016/s0925-2312(01)00676-2

[23] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520-8532, 2015. https://doi.org/10.1016/j.eswa.2015.07.007

[24] S. Barak, J. H. Dahooie, and T. Tichý, "Wrapper ANFIS-ICA method to do stock market timing and feature selection on the basis of Japanese Candlestick," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9221-9235, 2015. https://doi.org/10.1016/j.eswa.2015.08.010

[25] C.-H. Su and C.-H. Cheng, "A hybrid fuzzy time series model based on ANFIS and integrated nonlinear feature selection method for forecasting stock," *Neurocomputing*, vol. 205, pp. 264-273, 2016. https://doi.org/10.1016/j.neucom.2016.03.068

[26] S.-M. Chen, "Forecasting enrollments based on fuzzy time series," *Fuzzy Sets and Systems*, vol. 81, no. 3, pp. 311-319, 1996. https://doi.org/10.1016/0165-0114(95)00220-0

[27] A. Ben Ishak, "Variable selection using support vector regression and random forests: A comparative study," *Intelligent Data Analysis*, vol. 20, no. 1, pp. 83-104, 2016. https://doi.org/10.3233/ida-150795

[28] A. B. Ishak, M. B. Daoud, and A. Trabelsi, "Ozone concentration forecasting using statistical learning approaches," *J. Mater. Environ. Sci*, vol. 8, no. 12, pp. 4532-4543, 2017. https://doi.org/10.26872/jmes.2017.8.12.478

[29] M. Rana, M. M. Uddin, and M. M. Hoque, *Effects of activation functions and optimizers on stock price prediction using LSTM recurrent networks* Beijing, China: CSAI, 2019.

[30] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, "Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market," *IEEE Access*, vol. 8, pp. 22672-22685, 2020.

[31] A. U. Haq, A. Zeb, Z. Lei, and D. Zhang, "Forecasting daily stock trend using multi-filter feature selection and deep learning," *Expert Systems with Applications*, vol. 168, p. 114444, 2021. https://doi.org/10.1016/j.eswa.2020.114444

[32] L. Xie and S. Yu, "Unsupervised feature extraction with convolutional autoencoder with application to daily stock market prediction," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 16, p. e6282, 2021. https://doi.org/10.1002/cpe.6282

[33] S. Bhanja and A. Das, "A black swan event-based hybrid model for Indian stock markets' trends prediction," *Innovations in Systems and Software Engineering*, vol. 20, no. 2, pp. 121-135, 2024. https://doi.org/10.1007/s11334-021-00428-0

[34] V. Vapnik, "The nature of statistical learning theory," 1995.

[35] A. J. Smola, N. Murata, B. Schölkopf, and K. R. Müller, "Asymptotically optimal choice of ε-loss for support vector machines," in *Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN'98), vol. 1, pp. 105–110*, 1998.

[36] S. R. Gunn, M. Brown, and K. M. Bossle, "Network performance assessment for neuro-fuzzy data modelling. In Liu, X. Cohen P., & Berthold M. (Eds.)," *Intelligent Data Analysis, Lecture Notes in Computer Science*, vol. 1208, pp. 313–323, 1997. https://doi.org/10.1007/bfb0052850

[37] V. Vapnik, *Statistical learning theory*. New York: Springer, 1998.

[38] A. Rakotomamonjy, "Analysis of SVM regression bounds for variable ranking," *Neurocomputing*, vol. 70, no. 7-9, pp. 1489-1501, 2007. https://doi.org/10.1016/j.neucom.2006.03.016

[39] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. https://doi.org/10.1023/a:1010933404324

[40] L. Breiman, "Statistical modeling: The two cultures," *Quality Control and Applied Statistics*, vol. 48, no. 1, pp. 81-82, 2003.

[41] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157-1182, 2003. https://doi.org/10.1007/978-3-540-35488-8_1

[42] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389-422, 2002. https://doi.org/10.1023/a:1012487302797

[43] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83-90, 2006. https://doi.org/10.1016/j.chemolab.2006.01.007

[44] B. Ghattas and A. Ben Ishak, "Variable selection for high-dimensional binary classification: Comparisons and application to microarray data," *Journal de la Société Française de Statistique & Revue de Statistique Appliquée*, vol. 149, no. 3, pp. 43-66, 2008. https://doi.org/10.3917/rfs.491.0133

[45] C. Conrad, K. Loch, and D. Rittler, "On the macroeconomic determinants of long-term volatilities and correlations in US stock and crude oil markets," *Journal of Empirical Finance*, vol. 29, no. C, pp. 26-40, 2014. https://doi.org/10.1016/j.jempfin.2014.03.009

[46] E. Hajizadeh, A. Seifi, M. F. Zarandi, and I. Turksen, "A hybrid modeling approach for forecasting the volatility of S&P 500 index return," *Expert Systems with Applications*, vol. 39, no. 1, pp. 431-436, 2012. https://doi.org/10.1016/j.eswa.2011.07.033

[47] H. Kim and H. Park, "Term structure dynamics with macro-factors using high frequency data," *Journal of Empirical Finance*, vol. 22, pp. 78-93, 2013. https://doi.org/10.1016/j.jempfin.2013.03.003

[48] B. S. Paye, "'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables," *Journal of Financial Economics*, vol. 106, no. 3, pp. 527-546, 2012.

[49] S. Fang, M. Wang, W. Qi, and F. Zheng, "Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials," *Computational Materials Science*, vol. 44, no. 2, pp. 647-655, 2008. https://doi.org/10.1016/j.commatsci.2008.05.010

[50] M. Barua, T. Kumar, K. Raj, and A. M. Roy, "Comparative analysis of deep learning models for stock price prediction in the Indian market," *FinTech*, vol. 3, no. 4, pp. 551-568, 2024. https://doi.org/10.3390/fintech3040029

[51] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. Soman, "NSE stock market prediction using deep-learning models," *Procedia Computer Science*, vol. 132, pp. 1351-1362, 2018. https://doi.org/10.1016/j.procs.2018.05.050

Appendix A

S&P500 database

Variables	WHEAT	COFFEE	COPPER	OIL	GAZ	GOLD	EURUSD	USDJPY	GBPUSD
Code	1	2	3	4	5	6	7	8	9

O/C	H/C	L/C	Open	High	Low	NASDAQ	DJIA
10	11	12	13	14	15	16	17

S&P500 Ranking

Balpha	Bmargin	Bradius	Bspan	Salpha	Smargin	Sradius	Sspan	Srf
15	15	15	15	16	14	16	17	10
12	14	12	14	14	13	14	16	12
13	13	13	16	13	15	13	7	11
14	2	14	7	15	12	15	6	15
10	7	2	17	7	11	7	2	14
2	17	10	9	17	10	17	9	13
16	9	16	2	2	16	2	5	8
17	1	17	13	9	7	9	14	5
6	12	6	12	5	2	5	15	3
8	11	11	11	4	17	4	13	2
11	10	8	8	6	9	6	3	1
3	8	3	10	1	5	1	1	6
4	3	4	3	3	4	3	4	17
5	4	5	4	8	6	12	8	4
1	6	1	6	12	1	10	11	16
7	16	7	5	11	3	8	10	7
9	5	9	1	10	8	11	12	9

NASDAQ database

Variables	WHEAT	COFFEE	COPPER	OIL	GAZ	GOLD	EURUSD	USDJPY	GBPUSD
Code	1	2	3	4	5	6	7	8	9

O/C	H/C	L/C	Open	High	Low	S&P500	DJIA
10	11	12	13	14	15	16	17

NASDAQ Ranking

Balpha	Bmargin	Bradius	Bspan	Salpha	Smargin	Sradius	Sspan	Srf
14	15	14	15	16	14	16	17	10
11	14	11	14	14	13	14	7	12
13	13	15	13	13	15	13	16	11
10	17	13	2	15	11	15	9	15
15	2	2	17	7	12	7	5	14
2	7	10	16	9	10	9	3	13
12	9	17	1	17	16	17	2	3
17	16	12	7	5	7	5	4	17
8	12	8	12	6	17	6	1	8

6	11	6	11	3	9	3	6	1
5	10	5	10	1	5	1	14	5
4	8	4	8	2	6	2	15	16
16	3	9	5	4	3	4	13	2
9	4	16	3	8	2	8	8	9
3	6	3	6	12	1	12	11	7
7	1	7	4	11	4	10	12	6
1	5	1	9	10	8	11	10	4

DJIA database

Variables	**WHEAT**	**COFFEE**	**COPPER**	**OIL**	**GAZ**	**GOLD**	**EURUSD**	**USDJPY**	**GBPUSD**
Code	1	2	3	4	5	6	7	8	9

S&P500	**NASDAQ**	**O/C**	**H/C**	**L/C**	**Open**	**High**	**Low**
10	11	12	13	14	15	16	17

DJIA Ranking

Balpha	**Bmargin**	**Bradius**	**Bspan**	**Salpha**	**Smargin**	**Sradius**	**Sspan**	**Srf**
17	17	17	17	15	16	15	10	12
14	16	14	16	16	15	16	7	14
15	15	15	15	17	17	17	2	13
16	2	16	10	10	13	10	11	17
12	10	2	2	11	14	11	9	16
2	11	11	11	7	12	7	17	15
11	7	10	7	1	10	1	15	6
7	9	7	9	5	11	5	16	2
5	14	8	13	2	7	2	5	4
8	13	5	12	9	1	9	3	11
3	12	3	14	4	2	4	4	8
4	3	4	8	3	5	3	1	10
6	8	6	3	6	9	6	6	3
1	6	1	4	8	4	8	8	5
13	4	13	5	13	6	12	13	9
9	1	9	6	14	3	14	14	7
10	5	12	1	12	8	13	12	1